Internal Packing and Protein Structural Classes

J.W. Ponder and F.M. Richards
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06511

Protein folding is one of the most important and intriguing of the unsolved problems at the interface between chemistry and biology. The ability of polypeptide chains to fold into unique compact conformations with a vast array of biological functions is rare, if not unique, among high polymers. At this time there is no reason to believe that the chemical interactions currently recognized in small molecules are not directly applicable to proteins, and that the types of these interactions represent the full range that needs to be considered. Of course, the large size of macromolecules and the covalent connectivity will change the relative importance of the various interactions. Some effects barely detectable in small molecules may become dominant in large ones. Of particular interest in the protein folding problem is the observation that similar chain conformations may be seen for a whole series of amino acid sequences that may show little or no sequence homology. The central issue is: "What is the three-dimensional code?"

Drexler (1981) has suggested that it may be useful to express the folding problem in inverted form. Rather than ask the usual question, "What is the tertiary structure of a polypeptide chain of specified sequence?", ask the reverse, "What is the full list of sequences compatible with a given structure?" Pabo (1983) has pointed out that the elaboration of this suggestion may be useful for experiments in protein design as well as providing an approach to the folding problem. Although an answer to Drexler's second question would not remove the need for an answer to the first, it might provide some useful insights.

We are attempting to develop an algorithm that will provide a list of sequences that are compatible with a given structure. The list is called a tertiary template for the target structure. This phrase was introduced by Blundell and Sternberg (1985), and our use is a generalization of their definition. Our focus in the development of the algorithm has been a long-standing interest in intramolecular packing as a characteristic of proteins. Although this parameter can be evaluated in known structures (Richards 1974), its predictive use in the folding problem has remained elusive, even though it should play a major role. The use of packing considerations in the development of the templates is described below.

Background, Assumptions, and Structural Constraints

(1) The tertiary structures of known proteins appear to fall into general classes, as described by Levitt and Chothia (1976), Schulz and Schirmer (1979), and Richardson (1981). Although new classes will certainly be added to those presently known, we assume that the total number is finite and not very large. In recent years, the number of new classes has been rising much less rapidly than the number of new structures.

(2) We assume that the residues of any protein can be divided into two groups; those that are internal, i.e., only in contact with other protein atoms, and those that are external, i.e., partly or wholly in contact with solvent. We further assume that, for the purposes described below, precise division is not important. In general, there will be no simple relations between the sequence positions of either set of residues.

(3) We assume that the internal residues are responsible for the fold of the peptide chain and thus the protein class that it represents. Given an appropriate set of internal residues, the external residues affect the structure only permissively through the global free energy. The electrostatic or other properties of the external residues may affect the stability of the folded structure, but the geometry of the fold will be controlled entirely by the internal residues.

(4) Insertions and deletions in the peptide chains of different members of the class will be restricted to regions of the external set. The immediate expression of biological function will take place most obviously through the external residues that come into direct contact with the ligands. However, to the extent that global aspects of the structure, such as its dynamic behavior, are important, the internal residues will also be involved in function.

(5) Examination of known protein structures by many workers over a period of years has established the broad validity of the following general statements.

(a) The covalent geometry found in relevant small molecules may be used for proteins without significant change.

(b) As with all matter, atomic overlaps are prohibited. The best-known biochemical example is adherence of the actual structures to the Ramachandran map, which defines the allowed conformations of the peptide chain (Ramachandran and Sasisekharan 1968).

(c) Close packing results in small cavity volumes (Richards 1977).

(d) Buried hydrogen-bonding groups normally

[a] This paper is a modified version of Ponder and Richards (1987) with a few additional comments at the end.

occur as donor-acceptor pairs (Chothia 1976; Baker and Hubbard 1984).

(e) Groups with formal charges are located predominantly on the surface in contact with solvent (Janin 1979; Rashin and Honig 1984).

These appear to be strong statements that should severely restrict possible structures. However, it has been very difficult so far to apply these rules in a more than qualitative fashion.

The Procedure

The details of the algorithm in its current state of development are given by Ponder and Richards (1987). Only a brief outline is provided here.

(a) The *main chain* is considered to consist of the N, CA, C, O, and CB atoms. These are kept fixed in positions defined by the reference X-ray structure. The CB is removed only when it is in an interior position and is being tested as a possible Gly location.

(b) *Hydrogen atoms* are explicitly included on all carbon and nitrogen atoms in order to maximize the usefulness of the van der Waals overlap constraint in defining the packing. This has been commonly found necessary in other molecular packing studies.

(c) Ideally, the van der Waals contact and packing density portions of the algorithm would operate on all interior residues at once. This poses much too large a computational task. In practice, small interior volumes containing 5–8 residues are used in a single calculation and are referred to as *packing units*. The members of a unit are selected by visual inspection of the protein backbone structure on an interactive graphics terminal and are chosen in such a way that the side chains will point approximately toward the common centroid. Only then will the packing constraints described below operate efficiently. Sequence enumeration is performed for each of these packing units separately. This restriction is not as severe as it may first appear, since an overlapping group of such units may be processed and the results combined into one master template.

(d) The major computational difficulty of a packing study appears to be the conformational flexibility of the side chains. The number of angles involved is shown schematically in Figure 1. To address this problem we have reinvestigated the distribution of side-chain χ angles described some time ago by Janin et al. (1978). Using the 1985 version of the Protein Data Bank and selecting the most carefully refined high-resolution structures, we have found that the rotamer approximation for the angle distributions is much better now than in 1978 (see Fig. 2). The mean positions are very similar, but in all cases the standard deviations are markedly smaller. If we exclude for the moment Met, Lys, and Arg, all of the other 17 amino acids are adequately represented by fewer than 70 rotamers. These angles, and those represented by one standard deviation on either side, provide the *rotamer library*, and are used to represent the allowed conformations of the side chains.

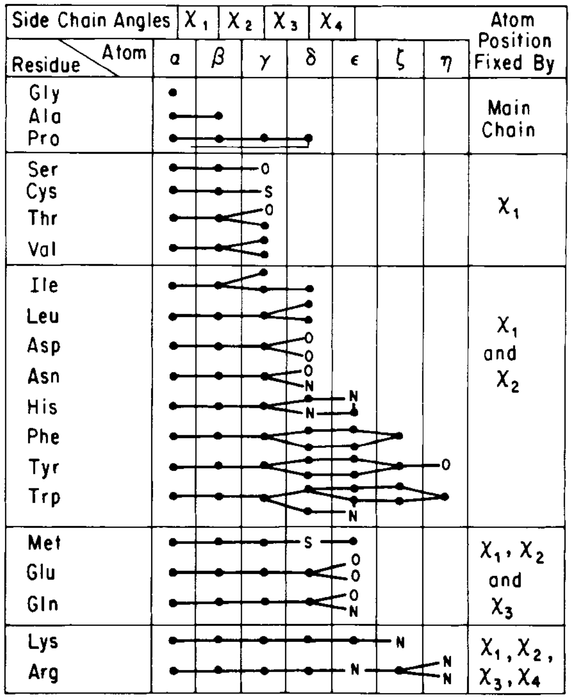

Figure 1. Flexibility of amino acid side chains. The figure shows the χ angle values required to fix the positions of side-chain atoms in each residue type. (Reprinted, with permission, from Ponder and Richards 1987.)

(e) The *van der Waals parameters* used are listed and discussed by Ponder and Richards (1987). Potential hydrogen-bonding groups were identified and treated separately.

(f) The high mean *packing density* found in all globular proteins is a strong structural constraint. The mean residue volumes found in a group of proteins have been calculated by Chothia (1975). Assigning these mean volumes to each residue, the total volume of the packing unit is taken as the sum of the individual residue volumes. We require that any proposed sequence fill the packing unit volume nearly as well as the native sequence. An input parameter allows adjustment of the precise percentage of native volume that will be considered acceptable.

(g) With the structure consisting solely of main-chain atoms (see a), the program goes sequentially through all the residues of the packing unit under study, the *main-chain/side-chain check*. At each position the full rotamer library is substituted one at a time and checked for steric overlap with other parts of the main chain. The restricted rotamer list for each position is filed for later use. Positions that must be occupied by Gly and those that must be Gly if another specified site is non-Gly are recorded. Potential hydrogen bonds to the main chain are also stored. Although there is wide variability in the number of rotamers permitted in the various positions, overall the restrictions introduced by the main chain reduce the number of candidate rotamer

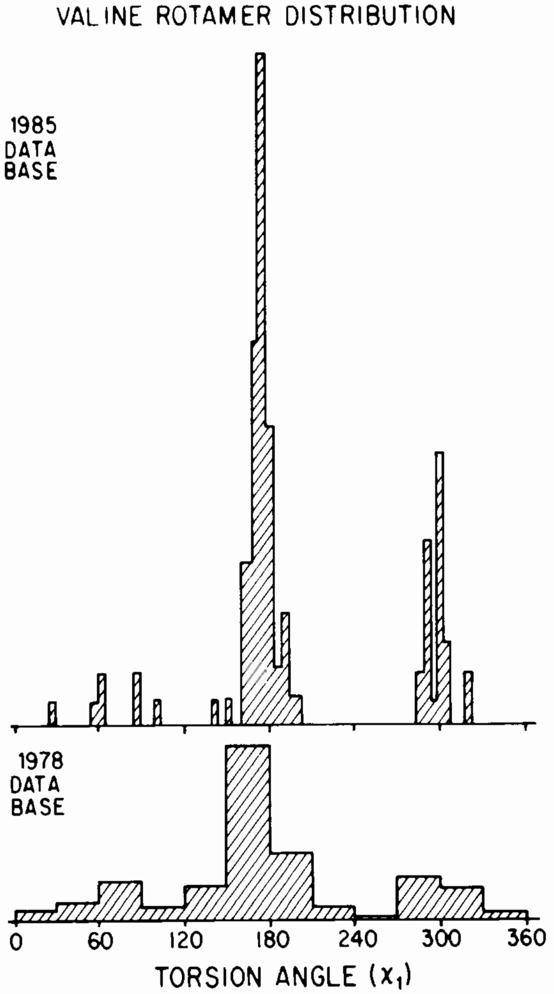

Figure 2. Comparison of 1985 and 1978 valine rotamer distributions. The upper section shows χ_1 values for the 151 Val residues in the 19-protein data base used in the current work. The lower section is a similar presentation for the 238 Val residues in the sample used by Janin et al. (1978). Sample sizes have been normalized so the total areas enclosed under each curve are equal. Similar narrowing of the 1985 distributions relative to 1978 values is observed for other amino acids. (Reprinted, with permission, from Ponder and Richards 1987.)

sequences by a factor of the order of 10^3–10^6. An additional option uses all atoms outside the packing unit as additional restricting atoms during the above checks. The implications of this option and its effects on results are discussed below.

(h) In the next step, all pairs of sites within the packing unit are surveyed, the *side-chain/side-chain check*. Starting with the restricted list from the main-chain check, all allowed rotamer pairs are determined and stored in a matrix. Possible side-chain/side-chain hydrogen bonds and potential disulfide bridges are noted as well. At this point, the time-consuming steric checks with all the required distance calculations are complete. In later testing for sequence acceptability for the packing unit, one only has to refer to this matrix for all pairs in the proposed sequence.

(i) The *combinatorial enumeration of allowed rotamer sequences* can now proceed. A proposed sequence is checked for steric overlap by table lookup. The packing density constraint is then introduced. A combinatorial tree search is applied in the generation of all nonredundant sequences. Both the allowed rotamer sequences and the corresponding decoded amino acid sequences are stored on disk files.

(j) For most proteins, a set of packing units will be required to cover all of the internal residues. If the units are picked to have overlapping residues, a mutually acceptable list of sequences can be selected from the lists for each unit computed separately. This final list is the *tertiary template*.

(k) Partial information from the full template can be presented in two-dimensional form, a *compositional template*, see Figure 3. The amino acid types are listed on the ordinate and the residue positions in the sequence on the abscissa. For a given sequence position,

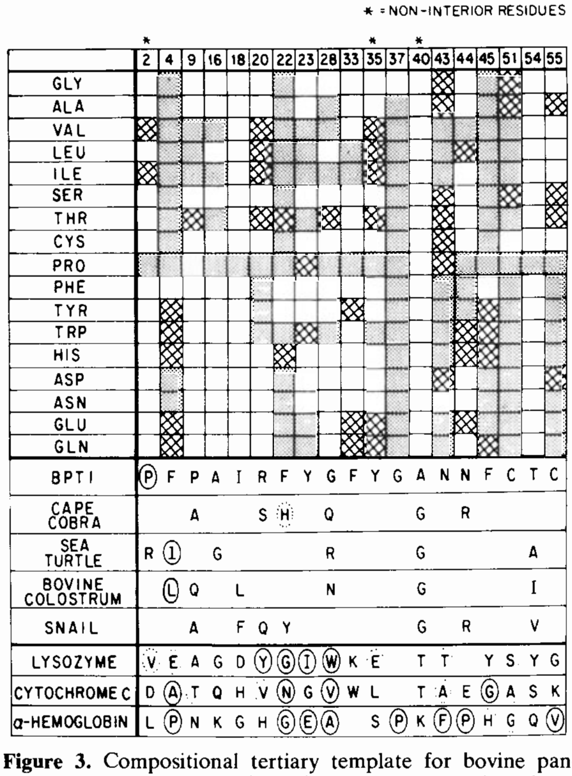
* = NON-INTERIOR RESIDUES
2 4 9 16 18 20 22 23 28 33 35 37 40 43 44 45 51 54 55
GLY
ALA
VAL
LEU
ILE
SER
THR
CYS
PRO
PHE
TYR
TRP
HIS
ASP
ASN
GLU
GLN
BPTI P F P A I R F Y G F Y G A N N F C T C
CAPE COBRA A S H Q G R
SEA TURTLE R I G R G A
BOVINE COLOSTRUM L Q L N G I
SNAIL A F Q Y G R V
LYSOZYME V E A G D Y G I W K E T T Y S Y G
CYTOCHROME C D A T Q H V N G V W L T A E G A S K
α-HEMOGLOBIN L P N K G H G E A S P K F P H G Q V

Figure 3. Compositional tertiary template for bovine pancreatic trypsin inhibitor (BPTI). The upper section shows permitted residue position/amino acid type combinations. Dotted combinations are forbidden in both the main-chain plus CB and the full-chain templates. Cross-hatched areas are allowed only in the main-chain plus CB template. Open areas indicate combinations allowed in both templates. The lower section shows sequence information for native BPTI, four homologous and three nonhomologous proteins. Sequence is provided only for those positions that differ from BPTI. Unbroken circles mark positions where the given sequence fails both templates. Dotted circles indicate failure with respect to only the full-chain template. In this illustration, all test sequences were tried only in the single alignment given by lining up the amino terminus of the test sequence with the amino terminus of BPTI. (Adapted from Ponder and Richards 1987.)

certain residues will not be found in any allowed sequence in the full template, and the appropriate element in the compositional template is blacked out. When complete, the clear positions in the two-dimensional template allow a rapid first estimate of the acceptability of a test sequence without computer intervention. Any sequence that will eventually pass the full template must also pass the compositional template. The reverse is not necessarily true. Many more sequences will pass the compositional test than will eventually pass the full template. Nonetheless, the compositional test may be useful and is easy to apply as a first screen.

Preliminary Results and Discussion

The rotamer library concept had to be tested first. Although the narrow distribution of χ values was encouraging, it was not clear how well the procedure would work for specific individual residues in the various structures. Individual interior residues in the native structures were removed, but all other atoms were left, and the full rotamer library was surveyed for replacements. In practically all cases in the most highly refined structures, and in more than 90% of the examples in the other structures, the native residue was found as the only acceptable residue, or among a small group considered acceptable, at each of the interior positions. The rotamer library approach thus appears to be a satisfactory approximation for this algorithm.

This observation by itself is of some interest. The mean χ angles were shown some time ago to correspond quite closely to the staggered conformations expected to represent energy minima. These positions are also those predicted on the basis of theoretical calculations on the isolated residues. There would thus appear to be no strain energy stored in the form of distorted torsional angles in the proteins of this basis set. Although an occasional residue may be forced into such a conformation, perhaps in the strong ligand fields found in metalloenzymes, for example, in general the residues will be in their most relaxed conformation in the native proteins. Since this is almost certainly true in the unfolded state, this observation may put substantial constraints on the paths for the folding reaction itself.

The next test was on three small proteins; crambin, rubredoxin, and scorpion neurotoxin. The packing units chosen for these proteins are shown in Figure 4. These proteins are so small, 46, 54, and 65 residues respectively, that they have very small "insides." However, there are some central residues in each protein, and the interior could be represented by a single packing unit. The tertiary templates were calculated for each protein and the statistics associated with these runs are shown in Table 1.

For all proteins the main-chain atoms plus CB represent about two-thirds of all the non-hydrogen atoms. Thus, the tremendous constraint on the allowed sequences produced simply by the main-chain overlap requirements is not unexpected. The increased restriction produced by the pairwise side-chain contacts is relatively small. This is due to the large number of small residue pairs where there will frequently be no side-chain contact at all. The packing constraint is again a very large factor, 10^2–10^4. The combined result of these constraints is an enormous reduction from the combinatorial maximum to the final number of allowed rotamer sequences, 100 to 300 in the examples chosen. There is a further reduction when these are reduced to the more conventional amino acid sequences. At this point, the only constraints are those of steric overlap and effective packing. If an additional constraint, such as absence of charged groups in the interior, is imposed, then the lists are further reduced.

The fact that a given amino acid sequence in many cases is represented by more than one rotamer sequence is interesting. This does not appear to correspond with the facts in known structures. The interior of most proteins is usually the clearest part of the X-ray structure with no evidence for multiple conformations of the side chains. In very high-resolution structures there are examples of multiple conformers, but these differ only slightly when found for interior residues (Smith et al. 1986). These observations show that even for the nonpolar residues, which make up the bulk of the tertiary templates, there are important energy terms not even implicitly accounted for in the present algorithm that select among the sterically acceptable rotamer sequences. An example of such a term would be the effect of the charge distribution in aromatic ring systems or carbonyl groups, as pointed out by Burley and Petsko (1985).

The enormous importance of the packing criterion in limiting the number of acceptable sequences is shown in Table 2. The first entry is the volume of the packing unit residues if they were all glycine. As the required volume for the packing unit is increased, there is at first very little effect, but soon there is a dramatic decrease in the number of allowed sequences. In the particular case of crambin, which is used for this example, the actual volume in the native structure is 659 Å^3. Note that there are no sequences that are capable of filling 700 Å^3 without steric overlap. There are possibly two sequences that might fill 690 Å^3. The difference between the native and the maximum packing is thus 30 Å^3. This volume is less than that of a single methyl group and only slightly larger than a methylene group. In this particular example, one cannot go from a sequence filling 660 Å^3 to one filling 690 Å^3 by simply changing one residue to another that is one carbon atom larger.

The rubredoxin case is interesting because of the large volume of the packing unit. Offhand, one might have thought that the larger volume would permit a much larger collection of acceptable sequences. This is not the case. The volume is so big that a large number of the largest residues are always required for proper filling, thus severely restricting the acceptable sequence list.

In recent computations, we have included in the

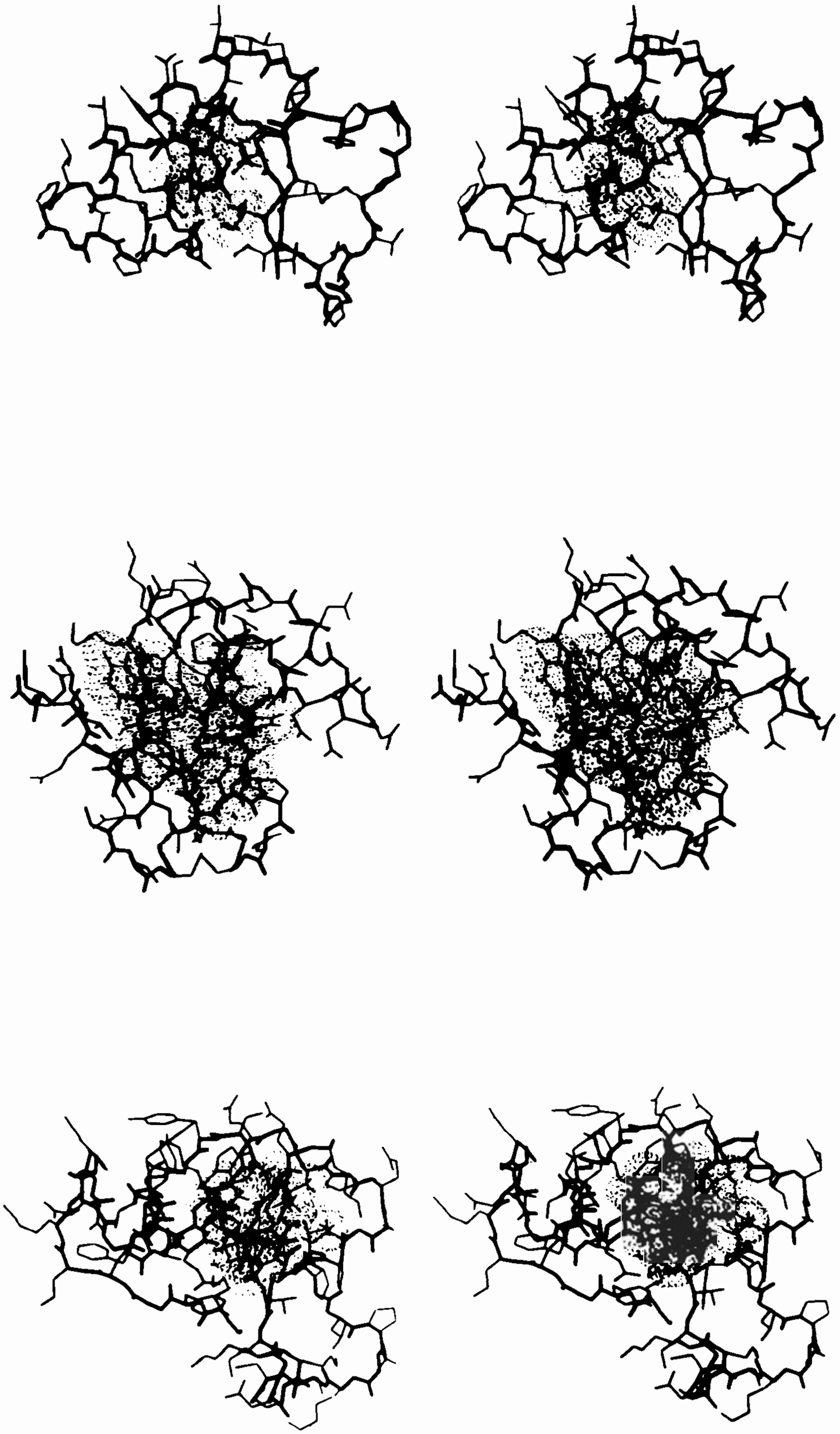

Figure 4. Stereo view of packing units in crambin, rubredoxin, and scorpion neurotoxin. Main-chain atoms are drawn in a heavy line. Hydrogen atoms are included only for residues contained in the packing unit. (*Upper*) Crambin with dot surface shown for residues 2, 13, 26, 30, and 32. (*Middle*) Rubredoxin with dot surface shown for residues 4, 11, 13, 30, 37, and 49. (*Lower*) Scorpion neurotoxin with dot surface shown for residues 5, 29, 34, 36, 48, 51, and 55. (Reprinted, with permission, from Ponder and Richards 1987.)

Table 1. Overall Results for Three Simple Tertiary Templates

	Number of sequences surviving sequential checks		
	crambin	rubredoxin	scorpion neurotoxin
Interior cavity:			
Number of residues	5	6	7
Volume in native	659	1256	897
Restricted by:			
None	1.4×10^9	9.1×10^{10}	6.1×10^{12}
Main chain only	85,652	152,915,040	4,717,440
Pairwise contacts	43,936	62,281,930	1,324,651
Packing constraint	95	236	284
Survivors as:			
Rotamer sequence	95	236	284
Amino acid sequence	34	44	70
Amino acid composition	18	20	58
Without charged residues:			
Amino acid sequence	34	30	14
Amino acid composition	18	14	13

(Reprinted, with permission, from Ponder and Richards 1987.)

rotamer library a tentative set of 13 Met structures. Interestingly, the sequence lists derived using the extended library often contain a large number of sequences with at least one Met residue. This observation is a direct result of the great flexibility of the Met side chain (i.e., three freely rotating χ angles). Conversely, a protein sequence containing a high proportion of Met sites might be able to arrange its side-chain conformations so as to fit several tyes of protein "fold." This could be one of many factors contributing to the relative rarity of the Met residue in known sequences.

Problems and Future Directions

If the template concept is to be useful, the sequence lists in the various templates must be mutually exclusive. As will be seen below, the present derivation is too restrictive to represent a class of structures rather than an individual structure. In addition, there will be errors in the reference structures and uncertainties in the proper allowed deviations. Thus, in checking a test sequence against a given template, it will be essential to derive a probability that the match is significant and not to demand that the match be perfect. Discrimination between templates will then be subject to the usual statistical tests. The appropriate weighting to set up such probabilities has yet to be worked out.

Table 2. Impact of Packing Constraint

Minimum volume (Å^3)	Rotamer sequences	Amino acid sequences	Amino acid compositions
330	43936	8113	1494
350	43935	8112	1493
400	43768	8043	1483
450	41877	7590	1442
500	31715	5867	1286
550	15188	3177	894
560	11922	2624	788
570	9041	2049	671
580	6469	1637	559
590	4633	1201	437
600	3288	874	338
610	1976	575	241
620	1340	398	168
630	803	236	102
640	464	137	65
650	199	74	35
660	93	33	17
670	43	17	9
680	9	5	4
690	2	1	1
700	0	0	0

(Reprinted, with permission, from Ponder and Richards 1987.)

Cavity Distribution

The marked dependence of the size of the template on the volume criterion assumed for a given packing unit is shown in Table 2. Although proteins are, on average, well packed, cavities or packing defects do exist. They represent a small fraction of the total volume but are not necessarily uniformly distributed. As long as the entire core of the protein may be treated as a single packing unit, the cavity problem is adequately handled by adjustment of the volume criterion. However, for larger cores, where current computing capacity requires division into multiple packing units, the cavity problem is more severe. The use of a single criterion for all units implies a uniform cavity distribution. If the criterion is relaxed to allow for an occasional large cavity, then the total cavity volume summed over all packing units becomes unrealistically large. The seriousness of this problem has yet to be evaluated in detail.

Interior and Exterior Residues

There is clearly an "interaction" between the interior and exterior residues. The packing and overlap criteria must eventually be applied to the interface between these two groups as well as to the interior residues alone. Thus, the largest list of template sequences will be those developed solely with restriction by main-chain atoms. The allowed sets of exterior residues, eventually put in to complete the protein, will depend on which of the template sequences is used for the interior set. In this indirect sense, surface portions of the structure directly involved in biological function may be affected by the internal set. In the reverse case, if one wishes to maintain the external set so as to ensure function, the template sequences will be substantially more restricted than they would be by just the main chain. The more restricted template may be the most useful in the design of experimental tests.

Insertions and Deletions

To compare a sequence with a template for a given class, one must be prepared to handle the problem of possible insertions and deletions. Our assumption is that all such changes occur in the external residues, that the interior core does not suffer insertions or deletions, and that the template is independent of such sequence changes. For this to work, the core residues must be arranged on elements of secondary structure that have a defined length and that are inviolate for a particular structural class. The template must include two or more sites on each element in order that the position and orientation of that element be fixed. Changes made in regions of the sequence between the elements are then irrelevant, since these positions are not used in defining the template. The relative positions of the template sites are specified, and the order of the elements in the full chain is known. Algorithms for such a search are similar to recently developed "template" methods for sequence homology searches in protein and nucleic acid data bases (see e.g., Gribskov et al. 1987).

Structure Variations within a Class

Even reduced solely to the template cores, the individual structures within a recognized class are not identical. The most serious problem for the present approach is represented by the structure variations within each class. There is at this time no clear definition of a class. The human eye has a marvelous ability to detect similarities in form, but the decision as to whether two objects are the "same" or not has a strong subjective component that is difficult to quantify. Chothia and Lesk (1986) have recently reported a relation between structural and sequence homology in several protein families. Their results imply that rms deviations of roughly 2.5 Å in main-chain atoms of core regions will accommodate sequences with essentially no homology. The computer modeling experiments by Novotný et al. (1984) on "misfolded" protein structures seem to substantiate this view. If the tertiary structure is allowed to vary, then inevitably the number of sequences in the template will increase, possibly dramatically. A logical procedure for expanding the template has not yet been found, and the extent of the required expansion is not yet known. To be useful, the templates will still have to be mutually exclusive.

The globin family makes a good test case. The "globin fold" is generally regarded as the same in myoglobin, the hemoglobins, and the plant protein leghemoglobin, even though between the distant relatives there is no recognizable sequence homology. In an interesting recent study, Elber and Karplus (1987) performed an extensive molecular dynamics simulation on myoglobin followed by energy minimization of selected frames. As part of this work, they noted that the relative conformations of the helices during the simulation sampled essentially all of the structures found in the globin class. Comparison of different minimized structures from the simulation showed main-chain rms deviations between frame pairs of as much as 2.7 Å. With the help of the above authors, we have started a sampling of the templates computed from selected frames in this simulation. To date, only one packing unit has been investigated, and this may be misleading. The X-ray structure is represented by a template of about 1100 amino acid sequences. For eight time frames picked to represent the most extreme structures, the total list increases to 15,000. Obviously, not all of these sequences are compatible with all of the structures. In fact, only 12 are found in common among the eight frames. The dynamic simulation is just that, and at any instant in time there may be steric overlap or large cavities that will subsequently disappear. The current algorithm is based on equilibrium structures and thus may not be appropriate for analyzing the dynamic simulations in its present form. Dynamic structures are only relevant to the template analysis insofar as they accurately represent the variation in equilibrium main-chain structure for members of the studied protein's class.

Experimental Tests

Even at the present stage of development, the templates provide an interesting goal for mutagenesis experiments. It is not obvious that evolution will necessarily have tested the full range of options for the interior residues for a given class of structures. Once a convenient core has been obtained, there would seem to be little pressure to change, since function is relegated to the external residue set, and this set, by definition, can be easily changed without affecting the core. Permissible single-site changes appear to be rare in the template sequences so far examined. Required multiple mutations would thus render interior residue changes even less likely.

There seems to be an admirable opportunity to test both the template proposal and its possible evolutionary significance by setting out to make the full set of permissible changes for at least one packing unit. This will require random mutations simultaneously at several sequentially distant sites. A strong selection procedure would seem to be mandatory. The experimental difficulties are formidable, but perhaps worth a try if the proper system can be found. Regardless of the fate of the current proposal, the results should be very interesting and useful. Some collaborative efforts along these lines have been initiated.

Other more selective tests may also be very useful. Specific mutations, which would be predicted either to permit or to prevent successful folding, could be made. Making the DNA for such a test would be more straightforward than a peptide synthesis approach. However, the expression of the gene may present a problem because isolation of the nonfolding mutants would be as important as those that fold successfully. Here, as well, the experimental problems are not simple. If a small enough system can be found, hopefully excluding disulfide bonds, organic synthesis of the peptides may be quite practical and free from some of the biological difficulties.

ACKNOWLEDGMENTS

F.M.R. especially thanks Carl Pabo for discussions in the early stages of this study. The work was supported by a grant from the National Institute of General Medical Sciences, GM-22778. J.W.P. was supported in part by a postdoctoral fellowship from the National Institute of General Medical Sciences, GM-11537.

REFERENCES

Baker, E.N. and R.E. Hubbard. 1984. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44:** 97.

Blundell, T. and M.J.E. Sternberg. 1985. Computer-aided design in protein engineering. *Trends Biotechnol.* **3:** 228.

Burley, S.K. and G.A. Petsko. 1985. Aromatic-aromatic interaction: A mechanism of protein structure stabilization. *Science* **229:** 23.

Chothia, C. 1975. Structural invariants in protein folding. *Nature* **254:** 304.

———. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105:** 1.

Chothia, C. and A.M. Lesk. 1986. The relation between divergence of sequence and structure in proteins. *EMBO J.* **5:** 823.

Drexler, K.E. 1981. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci.* **78:** 5275.

Elber, R. and M. Karplus. 1987. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. *Science* **235:** 318.

Gribskov, M., A.D. McLachlan, and D. Eisenberg. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci.* **84:** 4355.

Janin, J. 1979. Surface and inside volumes in globular proteins. *Nature* **277:** 491.

Janin, J., S. Wodak, M. Levitt, and B. Maigret. 1978. Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* **125:** 357.

Levitt, M. and C. Chothia. 1976. Structural patterns in globular proteins. *Nature* **261:** 552.

Novotný, J., R. Bruccoleri, and M. Karplus. 1984. An analysis of incorrectly folded protein models. *J. Mol. Biol.* **177:** 787.

Pabo, C. 1983. Designing proteins and peptides. *Nature* **301:** 200.

Ponder, J.W. and F.M. Richards. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193:** 775.

Ramachandran, G.N. and V. Sasisekharan. 1968. Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23:** 283.

Rashin, A.A. and B. Honig. 1984. On the environment of ionizable groups in globular proteins. *J. Mol. Biol.* **173:** 515.

Richards, F.M. 1974. The interpretation of protein structures: Total volume, group volume distributions and packing density. *J. Mol. Biol.* **82:** 1.

———. 1977. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* **6:** 151.

Richardson, J.S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34:** 167.

Schulz, G.E. and R.H. Schirmer. 1979. *Principles of protein structure*. Springer-Verlag, New York.

Smith, J.L., W.A. Hendrickson, R.B. Honzatko, and S. Sheriff. 1986. Structural heterogeneity in protein crystals. *Biochemistry* **25:** 5018.